Journal of Science, Innovation & Social Impact

2026 2nd International Forum on Computing and Intelligent Systems (IFCIS 2026)

Article

An Empirical Comparison of Multi-Agent LLM Collaboration Strategies for Cross-Document Fraud Evidence Aggregation and Auditable Investigation Chains

Yunfei Ga[1,*]

[1] Electrical Engineering, Northwestern University, IL, USA
* Correspondence: Yunfei Ga, Electrical Engineering, Northwestern University, IL, USA

Abstract: Suspicious-activity reporting under the Bank Secrecy Act requires investigators to aggregate evidence scattered across transaction records, customer identity files, behavioural logs, and external risk intelligence, while keeping every conclusion traceable to its source. Large language model agents have been proposed to support such cross-document investigation, yet little is known about how alternative collaboration strategies behave on this task. This work presents an empirical comparison of four established strategies --- single-agent serial reasoning, single-agent reasoning with self-consistency, role-specialised multi-agent collaboration, and debate-based multi-agent reasoning --- rather than proposing a new architecture. Using investigation cases synthesised from the AMLworld, Elliptic++, and Bank Account Fraud datasets, we evaluate each strategy along four axes: evidence recall, factual consistency, hallucination rate, and audit traceability. Across 300 cases on AMLworld, role-specialised collaboration attains the highest audit traceability (0.838), debate-based reasoning attains the highest factual consistency (0.841) and the lowest hallucination rate (0.094) at 6.7 times the cost of a single agent, and no strategy dominates on every axis. The gains are moderate and accompanied by substantial cost differences, indicating that strategy choice should be matched to the evidentiary and budgetary constraints of a given investigation.

Keywords: multi-agent reasoning; fraud investigation; evidence aggregation; auditable AI

Received: 29 April 2026
Revised: 17 June 2026
Accepted: 29 June 2026
Published: 03 July 2026

1. Introduction

1.1. Background and Motivation

Financial institutions in the United States must file suspicious-activity reports that document, with traceable evidence, why a transaction or account warrants regulatory attention. Compiling such a report is a cross-document task: an analyst reconciles raw transaction streams, know-your-customer identity files, longitudinal account-behaviour logs, and external risk intelligence before reaching a defensible conclusion. Much of this reconciliation remains manual, slow, and vulnerable to overlooked links, which has motivated interest in language-model agents that can read heterogeneous documents and assemble structured findings. A recent expert-guided multi-agent approach to cross-document fraudulent-evidence discovery illustrates both the promise and the difficulty of automating this pipeline, prioritising recall so that no material lead is missed [1].

Two properties separate fraud investigation from generic question answering. The first is verifiability: every assertion in a report must be anchored to a specific document passage, because an unsupported claim can invalidate a filing. The second is the cost of fabrication. Language models are prone to producing fluent but unsupported statements, a failure mode whose taxonomy distinguishes errors of factuality from errors of

faithfulness to the provided sources [2]. In a compliance setting, a faithfulness error --- a conclusion that contradicts or exceeds the evidence --- is more damaging than an incomplete answer, since it manufactures grounds for action that the underlying documents do not support.

These constraints make investigation a demanding testbed for agentic reasoning. A single agent reading every document in a single long context is simple to deploy yet offers little internal cross-checking, whereas multi-agent arrangements promise redundancy and explicit verification at the cost of higher latency and token consumption. Whether that price buys measurable gains in evidence recall, consistency, hallucination control, and traceability is an empirical question that has not been systematically examined for financial investigation.

1.2. Research Questions and Contributions

This work addresses that question through a controlled comparison rather than a new design.

1.3. Why Multi-Agent Reasoning for Fraud Investigation

The investigation pipeline decomposes naturally into retrieval, extraction, verification, and reporting, a structure that mirrors how multi-agent reasoning distributes sub-tasks across cooperating roles. Debate-style arrangements, in which several agents reason independently and then reconcile their conclusions, have been shown to improve factual accuracy on reasoning benchmarks by exposing and correcting individual errors [3]. The same mechanism is attractive when documents conflict --- for example, a sanctioned counterparty named in risk intelligence but absent from the identity file --- because disagreement among agents can surface the conflict rather than silently resolve it.

1.4. Scope of the Empirical Study

We compare four established strategies --- two single-agent baselines and two multi-agent collaboration designs --- holding the underlying reasoning engine fixed so that differences reflect coordination rather than base capability. Performance is measured along four axes relevant to compliance: evidence recall, factual consistency, hallucination rate, and audit traceability. The study deliberately avoids introducing a new architecture, complementing broad financial language-model benchmarks that report aggregate capability without isolating the effect of agent coordination on investigative quality [4]. Our contribution is the comparison itself and the evidence it provides about when coordination overhead is justified.

2. Related Work

Three lines of prior work inform this comparison: the coordination of multiple agents, the measurement of hallucination and faithfulness, and the detection of financial fraud over structured data.

2.1. LLM Multi-Agent Collaboration and Debate

2.1.1. Role-Specialised Collaboration

Conversational multi-agent programming abstracts an application into agents that exchange messages under configurable roles, letting a retriever, an extractor, and a writer cooperate without a monolithic prompt [5]. Encoding standard operating procedures as ordered role prompts further constrains the interaction, producing structured intermediate artefacts that reduce the cascading errors typical of free-form generation [6]. Carrying explicit message schemas between roles narrows the space of valid hand-offs, which matters when each piece of evidence must retain a reference to its origin. These designs motivate the role-specialised pipeline evaluated here, in which distinct agents handle evidence intake, evidence extraction, reasoning verification, and report generation.

2.1.2. Debate and Self-Consistency for Reasoning

A complementary line has agents reason independently and reconcile through discussion. Tit-for-tat debate managed by a judge agent counteracts the degeneration of

thought that afflicts single-agent self-reflection [7]. A lighter mechanism samples multiple reasoning paths and selects the most frequent answer, improving chain-of-thought accuracy without any inter-agent messaging [8]. Comprehensive agent evaluations show that the benefit of such coordination depends heavily on the base engine, with commercial models far outperforming open alternatives on long-horizon interactive tasks [9], which informs our decision to hold the engine fixed across strategies so that coordination is the only variable under study.

2.2. Hallucination and Factuality Evaluation

Measuring whether generated text is supported by sources is central to this study. Atomic-fact decomposition scores the factual precision of long-form generation by breaking a response into individual claims and verifying each against a knowledge source [10]. We adopt this decomposition to compute the hallucination and factual-consistency metrics, treating the synthesised documents as the knowledge source against which every generated proposition is checked. Grounding the metrics in claim-level verification keeps them aligned with the compliance requirement that each statement be defensible, rather than rewarding surface fluency that a coarser score might miss.

2.3. Financial Fraud Detection and Auditable Investigation

Most automated fraud detection operates on structured transactions rather than documents. Graph-based methods established the dominant paradigm, with the Elliptic dataset introducing labelled blockchain transactions for forensic analysis [11], and later work showing that adaptations which provably increase the expressive power of message passing on directed multigraphs improve detection of the illicit minority class in synthetic laundering data [12]. These detectors flag suspicious entities but stop short of producing the narrative- and source-anchored justification that a compliance filing requires. The gap between flagging an entity and assembling an auditable account of why it is suspicious is precisely what document-level reasoning aims to close, and it is the gap this comparison probes.

3. Experimental Setup

3.1. Task Formulation

We formalise an investigation case as a suspicious-activity alert paired with a set of associated documents drawn from four classes --- transaction streams, identity profiles, account-behaviour logs, and external risk intelligence. Given this input, a strategy must produce a structured report containing a disposition of escalate or dismiss, a set of supporting findings, and, for each finding, an explicit pointer to the document passage that substantiates it. The pointer requirement turns report generation into a grounded task: a finding without a valid passage reference is treated as unsupported, regardless of whether its content is correct, which allows the same output schema to feed into every evaluation metric without manual interpretation.

Because a single case may carry more documents than fit comfortably in a focused reasoning window, every strategy first selects the passages most relevant to the alert before reasoning over them. We implement this selection with dense passage retrieval, following the retrieval-augmented generation paradigm that pairs a parametric reader with a non-parametric document index [13]. Holding the retrieval component identical across strategies isolates the effect of how agents coordinate subsequently, so that observed differences in recall, consistency, hallucination, and traceability are attributable to the collaboration strategy rather than to retrieval quality. The retriever returns the same ranked passages to each strategy, and only the downstream reasoning differs.

3.2. Datasets and Document Synthesis

No public corpus pairs realistic suspicious-activity alerts with the heterogeneous supporting documents an investigator would consult, and document-level fraud-investigation benchmarks remain scarce. We derive investigation cases from three established structured datasets and render their records into documents. The primary

source is the AMLworld synthetic transaction collection, whose higher-illicit-ratio small split supplies roughly five million transactions across about 515 thousand accounts with complete laundering-pattern labels [14]. We draw cross-document cases from the dual-layer Elliptic++ dataset, which augments the transaction graph with 822 thousand wallet-level actors [15], and we use the Bank Account Fraud base variant to represent application fraud. Table 1 summarises the specifications of the three sources, and Figure 1 reports the composition of the resulting cases (As shown in Table 2).

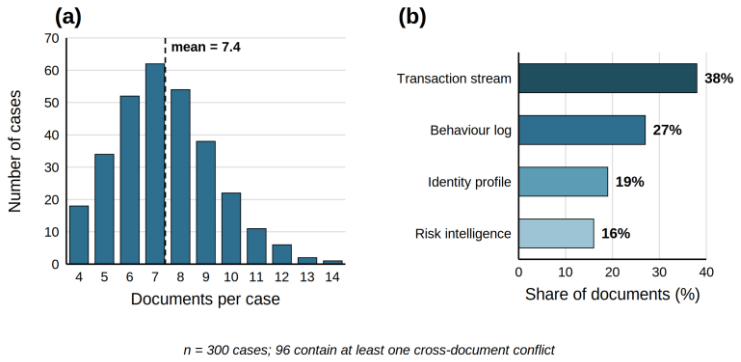

Figure 1. Composition of Synthesised Investigation Cases

Distribution of supporting documents across the 300 synthesised investigation cases. (a) Number of associated documents per case, averaging 7.4 with a range of 4 to 14. (b) Composition by document class, with transaction-stream entries the most frequent at 38%, followed by account-behaviour logs at 27%, identity profiles at 19%, and external risk-intelligence notes at 16%. Cases containing at least one cross-document conflict account for 96 of the 300.

Table 1. Specifications of the Source Datasets

Dataset	Source / reference	Scale	Labels
AMLworld HI-Small	IBM Research; Altman et al., NeurIPS 2023	5M transactions, 515K accounts, ~0.10% illicit	Complete laundering-pattern labels
Elliptic++	Georgia Tech; Elmougy & Liu, KDD 2023	203K transaction nodes, 822K wallet actors (56 features), 1.27M interactions, 49 time steps	Licit / illicit / unknown
Bank Account Fraud (Base)	Feedzai, NeurIPS 2022	1,000,000 records, 32 features, ~1% fraud	Binary application-fraud label

Table 2. Configuration of the Compared Collaboration Strategies

Strategy	Agents	Coordination	Decoding	Rel. cost
Single-Agent Serial	1	None (single pass)	Greedy	1.0×
Single-Agent + Self-Consistency	1	Majority vote over 5 samples	Sampled	4.8×

Role-Specialised	4	Structured message passing	Greedy	3.6×
Debate-Based	3 + 1 judge	2 rounds of critique, then consolidation	Sampled	6.7×

Specifications drawn from the respective dataset documentation and publications.

3.2.1. Tabular-to-Document Transformation

Each structured record is rendered into natural-language documents through templated prompts that preserve every source field value while adding the surface form an investigator would encounter. A transaction row becomes a transaction-stream entry narrating the date, counterparties, amount, currency, and channel; the same account's aggregated history becomes a behaviour log summarising ninety-day activity, login devices, and geographic patterns; account metadata becomes an identity profile; and a synthetic risk-intelligence note records sanction-list matches and adverse-media mentions. Attributes that the structured sources do not supply --- the beneficial owner, industry, and registered jurisdiction on the identity profile, together with the sanction and adverse-media entries on the risk-intelligence note --- are filled deterministically from a fixed schema keyed to the record and its label, for example by attaching a sanction match to an account already labelled illicit, so that every such attribute is reproducible from a known generating rule rather than invented anew for each document. To prevent the rendering step from injecting unsupported content, a verification pass checks every source-derived figure in each generated document against the originating record and discards any document that introduces a value neither carried over from the source nor produced by this schema. The ground-truth evidence for each case is therefore the set of passages traceable either to a labelled record or to the deterministic schema applied to it, which is the reference against which the evaluation metrics are computed.

3.2.2. Suspicious Alert Case Construction

Cases are built around labelled events so that the ground truth is unambiguous. For a flagged account or transaction, we assemble the alert together with the documents generated from its one- and two-hop neighbourhood, mixing genuinely illicit cases with plausible false positives in a ratio that reflects the heavy class imbalance of real screening. A subset of cases is constructed to contain cross-document conflicts, in which a counterparty flagged in risk intelligence is missing from the identity file or a behaviour log contradicts a stated transaction purpose, since these are the situations where coordinated cross-checking should matter most. The evaluation set comprises 300 cases on the primary source, of which 96 contain at least one cross-document conflict. Figure 1 shows that cases carry 7.4 associated documents on average, dominated by transaction-stream entries at 38% and behaviour logs at 27%, with identity profiles at 19% and risk-intelligence notes at 16% completing each dossier.

3.3. Collaboration Strategies under Comparison

Four strategies are compared under an identical reasoning engine, retrieval component, and document set, with GPT-4o serving as the base reasoning engine for every agent. Table 2 lists their configurations, including the relative inference cost that later analysis revisits.

3.3.1. Single-Agent Serial Reasoning Baseline

The baseline routes all retrieved passages to one agent, which reads them sequentially and emits the structured report in a single pass. This arrangement carries no coordination overhead and serves as the reference point against which multi-agent gains are measured. A strengthened variant applies self-consistency by sampling five independent reasoning paths at a higher decoding temperature and retaining the

disposition and findings that recur across a majority of samples, trading additional computation for resistance to any single unstable generation. Both variants share the same prompt and output schema as the multi-agent strategies, differing only in the absence of inter-agent communication, which keeps the baseline a clean control.

3.3.2. Role-Specialised and Debate-Based Pipelines

The role-specialised pipeline assigns four agents --- an evidence intake agent, an evidence extractor, a reasoning verifier, and a report generator --- that pass structured messages carrying candidate evidence, each annotated with its source passage identifier. The intake agent performs no retrieval of its own; it receives the shared ranked passages produced by the common retriever and organises them for the extractor, so that the fixed retrieval component remains identical across strategies and only the downstream coordination differs. The role-playing prompt design that elicits cooperative behaviour between paired agents informs how these roles are instructed [16], and the verifier step adopts the communicative cross-checking shown to suppress unsupported claims during multi-agent generation. The debate-based strategy instantiates three reasoning agents that independently analyse the same evidence set and then exchange critiques over two rounds, after which a separate judge agent consolidates the surviving findings by majority agreement. Both pipelines emit the identical report schema, so the comparison reflects how evidence is aggregated and checked rather than differences in output format.

3.4. Implementation Details and Evaluation Metrics

Every agent is driven by the GPT-4o-2024-08-06 snapshot via the chat-completion interface, with generation capped at 2,048 tokens per call. The single-agent serial baseline and the role-specialised pipeline decode greedily at temperature zero, while the self-consistency and debate strategies sample at temperature 0.7 with nucleus sampling at top-p 0.95; the self-consistency variant draws five samples and resolves them by majority vote, and the debate strategy runs three reasoning agents over two critique rounds before a separate judge agent consolidates the outcome. Passage selection is shared by every strategy: a dense dual-encoder retriever following the formulation of [17] indexes each case in passages of roughly two hundred words and returns the eight highest-scoring passages per alert, so that any difference in performance reflects coordination rather than retrieval. Every strategy receives the same base investigation instructions, retrieved passages, passage identifiers, and output schema, while role-specialised agents receive additional role-specific instructions for intake, extraction, verification, or report generation.

The four quality axes are scored at the level of individual claims rather than whole reports, so that a single unsupported proposition is penalised even when the surrounding report is otherwise sound. Each report is decomposed into atomic findings following [18]; the ground-truth evidence of a case is the set of source-traceable passages that substantiate its label; and a finding is counted as supported only when the passage it cites both exists and entails the claim. Evidence recall is the number of ground-truth passages cited in the report divided by the total number of ground-truth passages, measuring how much of the material a strategy recovers. Factual consistency is the number of findings judged to be supported or substantially entailed by the cited evidence divided by the total number of generated findings, averaged over reports. Hallucination rate is computed separately as the share of generated findings that introduce material entities, events, relationships, or risk conclusions not supported by any retrieved passage; partially supported or ambiguous findings are penalised in factual consistency but are not automatically counted as hallucinations. Audit traceability measures the share of generated findings that preserve a valid, correctly formatted, and reviewable source pointer through the final report. A pointer is counted as traceable when it refers to an existing retrieved passage and provides a reviewer with a clear path back to the cited evidence; this metric therefore evaluates citation completeness and auditability, while factual consistency separately assesses whether the claim itself is fully supported by the evidence.

Claim-level support and entailment are adjudicated by a judge agent built on the same GPT-4o-2024-08-06 snapshot but instantiated separately from the generating agents and prompted to mitigate the position and verbosity biases documented for language-model evaluators [19]. The scoring is otherwise fully automated; this study does not enlist a human compliance reviewer, and the consequences of relying on an automated judge are discussed in the limitations section (Section 5.2).

4. Results and Analysis

4.1. Aggregate Comparison Across Strategies

We report mean performance over five random seeds on the 300-case AMLworld evaluation set. Table 3 presents the four quality metrics together with inference cost relative to the single-agent baseline, and Figure 2 visualises the same comparison.

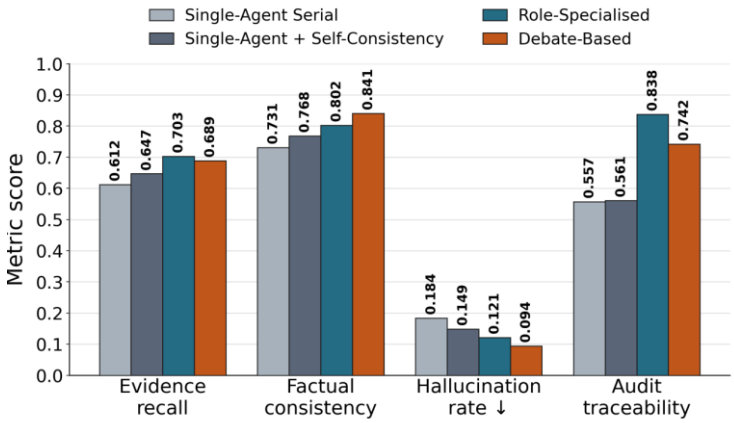

Figure 2. Quality Metrics Across Collaboration Strategies

Comparison of the four collaboration strategies on the 300-case AMLworld evaluation set along evidence recall, factual consistency, hallucination rate, and audit traceability. Role-specialised collaboration attains the highest evidence recall at 0.703 and audit traceability at 0.838, while debate-based reasoning attains the highest factual consistency at 0.841 and the lowest hallucination rate at 0.094. The single-agent baseline trials on every quality axis, and the self-consistency variant occupies an intermediate position. No strategy leads on all four metrics.

Table 3. Quality Metrics and Inference Cost on the AMLworld Evaluation Set

Strategy	Evidence recall	Factual consistency	Hallucination rate ↓	Audit traceability	Rel. cost
Single-Agent Serial	0.612	0.731	0.184	0.557	1.0×
Single-Agent + Self-Consistency	0.647	0.768	0.149	0.561	4.8×
Role-Specialised	0.703	0.802	0.121	0.838	3.6×
Debate-Based	0.689	0.841	0.094	0.742	6.7×

Mean over five seeds on 300 cases; the best values are split between the role-specialised and debate-based strategies. Source: experiments described in this work.

4.1.1. Evidence Recall and Factual Consistency

Role-specialised collaboration achieves the highest evidence recall at 0.703, recovering roughly 9 percentage points more of ground-truth evidence than the single-agent baseline at 0.612, an improvement attributable to the dedicated extraction role that examines each passage in isolation. Debate-based reasoning recovers slightly less evidence at 0.689 yet attains the highest factual consistency at 0.841, against 0.731 for the single agent, indicating that reconciling independent analyses removes propositions that a single pass would have asserted without sufficient support. The self-consistency variant lands between the baseline and the multi-agent strategies on both metrics, with 0.647 recall and 0.768 consistency, confirming that majority voting recovers some, but not all, of the benefit of explicit role separation or debate. The multi-role evaluation arrangement that improves alignment with human judgement [18] is mirrored here, where distributing analysis across agents yields more reliable consistency than any single perspective.

4.1.2. Hallucination Rate and Audit Traceability

Debate-based reasoning produces the lowest hallucination rate at 0.094, roughly half the 0.184 of the single-agent baseline, while role-specialised collaboration reaches 0.121. The pattern reverses on audit traceability, where the role-specialised pipeline leads decisively at 0.838 against 0.557 for the single agent, because its messages carry source-passage identifiers from extraction through to the final report, while the debate strategy reconciles conclusions without always preserving the originating reference and reaches 0.742. The self-consistency variant barely improves traceability over the baseline, from 0.557 to 0.561, since voting changes which findings survive but not whether each is anchored to a passage. Automated scoring was performed by a judge agent prompted to mitigate the position and verbosity biases documented for language-model evaluators [19], and the resulting metrics show that no single strategy is best on every axis, with recall and traceability favouring role specialisation while consistency and hallucination control favour debate.

4.2. Performance on Cross-Document Conflict Cases

The advantage of coordinated reasoning is clearest on the 96 cases containing cross-document conflicts, where evidence must be reconciled rather than merely collected. Table 4 reports the factual consistency and hallucination rates for this subset. Every strategy scores below its full-set average here, confirming that conflicting documents are genuinely harder, yet the gap between debate and the single agent widens substantially. The single-agent baseline hallucinates on 0.241 of conflict-case propositions, while debate reduces this to 0.112, a drop of nearly thirteen percentage points that exceeds the nine-point reduction observed on the full set. Factual consistency on conflict cases follows the same ordering, rising from 0.662 for the single agent to 0.808 for debate, with role specialisation intermediate at 0.749. This widening gap supports the intuition that independent re-analysis and cross-examination are most valuable precisely when documents disagree, since a single agent reading conflicting sources in a single pass tends to silently favour one account, whereas debate surfaces the contradiction and forces it to be resolved against the evidence.

Table 4. Factual Consistency and Hallucination Rate on Cross-Document Conflict Cases

Strategy	Factual consistency	Hallucination rate ↓
Single-Agent Serial	0.662	0.241
Single-Agent + Self-Consistency	0.704	0.198
Role-Specialised	0.749	0.163
Debate-Based	0.808	0.112

Subset of 96 conflict cases from the AMLworld evaluation set. Source: experiments described in this work.

4.3. Cross-Dataset and Cost Analysis

To test whether these patterns are specific to a single source, we repeat the comparison using Elliptic++ and the Bank Account Fraud base variant. Table 5 reports evidence recall and hallucination rates for the three primary strategies --- the single-agent serial baseline, role specialisation, and debate --- setting aside the self-consistency variant --- across all three datasets, and Figure 3 relates hallucination rate to inference cost.

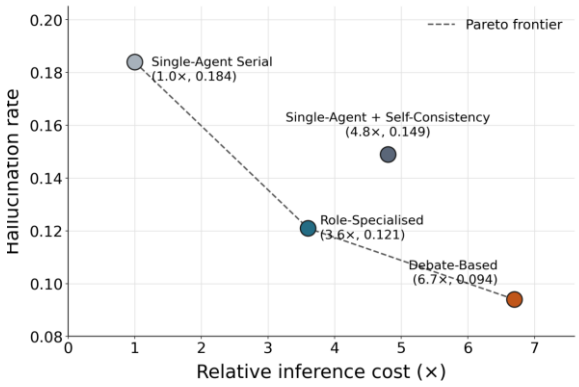

Figure 3. Hallucination Rate Against Relative Inference Cost

Relationship between hallucination rate and inference cost relative to the single-agent baseline. The single-agent baseline sits at unit cost with a hallucination rate of 0.184; the role-specialised pipeline at 3.6 times cost with 0.121; the self-consistency variant at 4.8 times cost with 0.149; and debate-based reasoning at 6.7 times cost with 0.094. Debate reaches the lowest hallucination rate at the highest cost, while role specialisation offers the largest hallucination reduction per unit of additional cost.

Table 5. Cross-Dataset Evidence Recall and Hallucination Rate

Dataset	Metric	Single-Agent	Role-Specialised	Debate-Based
AMLworld HI-Small	Evidence recall	0.612	0.703	0.689
AMLworld HI-Small	Hallucination rate ↓	0.184	0.121	0.094
Elliptic++	Evidence recall	0.578	0.664	0.651
Elliptic++	Hallucination rate ↓	0.203	0.139	0.107
Bank Account Fraud	Evidence recall	0.634	0.719	0.706
Bank Account Fraud	Hallucination rate ↓	0.171	0.114	0.089

Source: experiments described in this work.

4.3.1. Generalisation Across Datasets

The relative ordering of the strategies is stable across datasets. Role-specialised collaboration leads in evidence recall across all sources, at 0.703 on AMLworld, 0.664 on Elliptic++, and 0.719 on Bank Account Fraud, while debate-based reasoning attains the lowest hallucination rate across all settings, ranging from 0.089 to 0.107. Absolute scores are weakest on Elliptic++, where the dual-layer transaction-and-wallet structure produces denser document sets and more opportunities for spurious association, and strongest on Bank Account Fraud, whose application records yield more self-contained cases. The

persistence of the ordering across sources of differing structure echoes findings from graph-based fraud detection, where methods that explicitly handle adversarial camouflage retain their relative advantage across settings [20]. The stability suggests that the trade-off between traceability-oriented role specialisation and consistency-oriented debate is a property of the collaboration strategy rather than an artefact of any single dataset.

4.3.2. Cost-Effectiveness Trade-Offs

Quality gains carry markedly different costs. The single-agent baseline runs at unit cost; role specialisation at 3.6 times that cost; self-consistency at 4.8 times that cost; and debate at 6.7 times that cost, as shown in Figure 3. Debate buys the lowest hallucination rate but at the highest price, while role specialisation offers the most favourable reduction in hallucination per unit of additional cost and delivers the best traceability. Relative to the single-agent baseline, role specialisation reduces hallucination by 0.063 at an additional cost of 2.6×, or about 0.024 per extra cost unit, whereas debate reduces hallucination by 0.090 at an additional cost of 5.7×, or about 0.016 per extra cost unit. On cases that hinge on a single decisive document, the four strategies converge to within a few points on every quality metric, leaving the unit-cost single-agent baseline the most economical choice and reserving the multi-agent strategies for dossiers where evidence is genuinely distributed. These differences are moderate in magnitude --- the largest quality gap on any single axis is the traceability advantage of role specialisation --- and they do not point to one universally preferable strategy, but to a mapping from investigative conditions and budget onto an appropriate level of coordination.

5. Discussion and Future Work

5.1. Discussion

The comparison yields a consistent picture: coordination among agents improves the qualities a compliance investigation cares about, but the improvements are partitioned rather than uniform. Role specialisation, by threading source-passage identifiers through every message, makes the resulting report auditable in a way a single agent does not, which matters when each conclusion must withstand regulatory scrutiny. Debate, by forcing independent analyses to confront one another, suppresses unsupported claims and raises consistency, an advantage that grows when documents conflict. That these strengths are complementary rather than coincident is the central practical finding, since it implies that an operational pipeline need not choose a single strategy outright but can route cases according to their evidentiary profiles. A dossier whose verdict turns on reconciling contradictory sources benefits most from debate, while a filing that will be reviewed line by line benefits most from the traceability of role specialisation. The single-agent baseline retains a clear role for the large volume of cases that resolve on a single document, where the overhead of coordination earns little. Framing the result this way shifts the deployment question away from searching for the best agent arrangement and toward a triage policy that matches coordination costs to case difficulty, which is a more actionable conclusion for institutions operating under both accuracy and budget constraints.

Limitations: Several limitations qualify these findings. The investigation cases are synthesised from structured datasets rather than drawn from authentic suspicious-activity files, so the rendered documents, while internally consistent, may understate the noise, ambiguity, and adversarial obfuscation of real dossiers. The absence of public document-level investigation corpora makes this gap difficult to close at present. All strategies share a single reasoning engine, which removes base capability as a confound but leaves open whether the observed ordering persists under weaker or differently trained engines. The automated scoring, although calibrated against bias-aware prompting, inherits whatever residual preferences the judge retains. The cost figures are relative ratios measured under a single configuration and will vary with context length, retrieval depth, and the number of debate rounds. Future work should validate the strategies on authentic, access-controlled investigation records under appropriate governance, introduce a small panel of human

compliance reviewers to corroborate the automated metrics, and test mixed strategies that invoke debate only for cases flagged as conflicting while handling the remainder with the cheaper baseline. Coupling the document-level reasoning examined here with the structured detectors that flag suspicious entities in the first place would also close the loop from alert to auditable report, which is the practical objective these strategies are meant to serve.

References

1. S. Bai, B. Wu, Y. Zhang, C. Wu, X. Zheng, Y. Yuan, K. Wu, and J. Li, "AuditAgent: Expert-guided multi-agent reasoning for cross-document fraudulent evidence discovery," in *Proc. 6th ACM Int. Conf. AI in Finance (ICAIF '25)*. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://arxiv.org/abs/2510.00156
2. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Art. no. 42, 2025. doi: 10.1145/3703155
3. Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proc. 41st Int. Conf. Mach. Learn. (ICML 2024)*. Proceedings of Machine Learning Research, 2024. [Online]. Available: https://arxiv.org/abs/2305.14325
4. Q. Xie et al., "FinBen: A holistic financial benchmark for large language models," in *Adv. Neural Inf. Process. Syst. 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024. [Online]. Available: https://arxiv.org/abs/2402.12659
5. Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," arXiv preprint arXiv:2308.08155, 2023. [Online]. Available: https://arxiv.org/abs/2308.08155
6. S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber, "MetaGPT: Meta programming for a multi-agent collaborative framework," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR 2024)*, 2024. [Online]. Available: https://arxiv.org/abs/2308.00352
7. T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, "Encouraging divergent thinking in large language models through multi-agent debate," in *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2024)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 17889–17904. [Online]. Available: https://arxiv.org/abs/2305.19118
8. X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR 2023)*, 2023. [Online]. Available: https://arxiv.org/abs/2203.11171
9. X. Liu et al., "AgentBench: Evaluating LLMs as agents," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR 2024)*, 2024. [Online]. Available: https://arxiv.org/abs/2308.03688
10. S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proc. 2023 Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2023)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 12076–12100. [Online]. Available: https://arxiv.org/abs/2305.14251
11. M. Weber, G. Domeniconi, J. Chen, D. K. I. Weidele, C. Bellei, T. Robinson, and C. E. Leiserson, "Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics," in *KDD '19 Workshop Anomaly Detect. Finance*, 2019. [Online]. Available: https://arxiv.org/abs/1908.02591
12. B. Egressy, L. von Niederhäusern, J. Blanuša, E. Altman, R. Wattenhofer, and K. Atasu, "Provably powerful graph neural networks for directed multigraphs," in *Proc. 38th AAAI Conf. Artif. Intell. (AAAI-24)*, vol. 38, no. 10, 2024, pp. 11838–11846. doi: 10.1609/aaai.v38i10.29069
13. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. 33 (NeurIPS 2020)*, 2020, pp. 9459–9474. [Online]. Available: https://arxiv.org/abs/2005.11401
14. E. Altman, J. Blanuša, L. von Niederhäusern, B. Egressy, A. Anghel, and K. Atasu, "Realistic synthetic financial transactions for anti-money laundering models," in *Adv. Neural Inf. Process. Syst. 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. [Online]. Available: https://arxiv.org/abs/2306.16424
15. Y. Elmougy and L. Liu, "Demystifying fraudulent transactions and illicit nodes in the Bitcoin network for financial forensics," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discov. Data Mining (KDD '23)*. New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3580305.3599803
16. G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: Communicative agents for 'mind' exploration of large language model society," in *Adv. Neural Inf. Process. Syst. 36 (NeurIPS 2023)*, 2023, pp. 51991–52008. [Online]. Available: https://arxiv.org/abs/2303.17760
17. C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, "ChatDev: Communicative agents for software development," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL 2024)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 15174–15186. [Online]. Available: https://arxiv.org/abs/2307.07924

18. C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "ChatEval: Towards better LLM-based evaluators through multi-agent debate," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR 2024)*, 2024. [Online]. Available: https://arxiv.org/abs/2308.07201
19. L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Adv. Neural Inf. Process. Syst. 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. [Online]. Available: https://arxiv.org/abs/2306.05685
20. Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, "Enhancing graph neural network-based fraud detectors against camouflaged fraudsters," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage. (CIKM '20)*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 315–324. doi: 10.1145/3340531.3411903